Estimation of Viral Population Structure from Amplicon-Based Reads

Nicholas Mancuso,
Alexander Artyomenko, Alex Zelikovsky
Department of Computer Science
Georgia State University
Atlanta, Georgia 30302–3994
Email: {nmancuso, aartyomenko, alexz}@cs.gsu.edu

Pavel Skums
Centers for Disease Control
and Prevention
Atlanta, Georgia 30333
Email: kki8@cdc.gov

Ion Măndoiu
Department of Computer Science
& Engineering
University of Connecticut
Storrs, CT 06269
Email: ion@engr.uconn.edu

***Abstract*—Accurately estimating the structure of highly diverse viral populations is a challenging task. There are two main impediments to globally reconstructing a population. The first is the presence of sequencing errors in reads. Judiciously differentiating these errors from actual rare variants must be properly handled or the global structure may be ill-defined. Secondly, long conserved regions in the viral genome extend beyond what modern sequencers are capable of producing. As a result, the actual population diversity may be hidden in these targeted regions. We propose VirA, a tool for global reconstruction of a viral population that overcomes these obstacles by combining local error correction and a read-graph approach.**

***Keywords*—next-generation sequencing; viral quasispecies; global reconstruction;**

I. Discussion

A *quasispecies* is a heterogeneous closely-related intra-host viral population. Estimating the structure (i.e., variants and their respective abundance) is categorized into two approaches: *local* and *global* reconstruction. Local methods focus on a single targeted PCR region and estimate the diversity. Global reconstruction methods target longer regions and assemble the population.

One approach to sequencing viral populations is to perform PCR for multiple overlapping regions. The resulting pools are then combined and sequenced by technologies such as Roche 454 or Ion Torrent. VirA takes as input the sequenced reads, a reference genome, and the target-specific primers used for PCR. It estimates the viral population by performing the following steps:

1) align the reads to the reference[1]
2) identify each amplicon (i.e., targeted region) and locally reconstruct the population via kGEM[2]
3) construct a read-overlap graph
4) assemble the population using either a Maximum Bandwidth[3] or Multi-commodity Flow method[4].

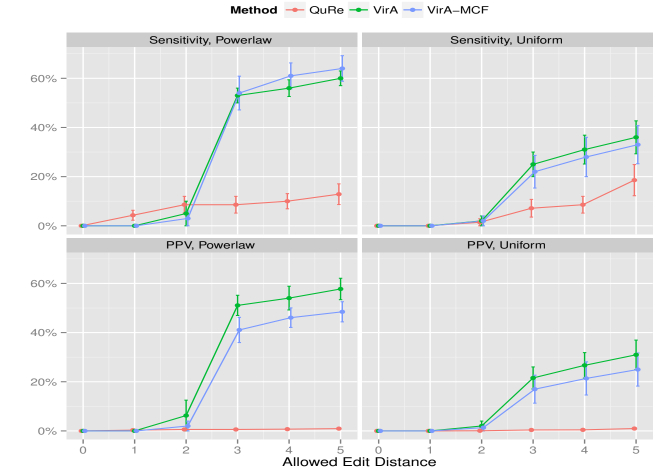

Fig. 1. Sensitivity and PPV on Simulated Data

Preliminary results on 20 simulated HCV E1E2 data show VirA to outperform QuRe in terms of quality of the solution (Figure 1). Each dataset contained 8-12 amplicon (20k reads each) regions spanning 1734bp over 10 variants. All subpopulations conformed to either a uniform or powerlaw (with $\alpha = 2.0$) distribution. Reads were generated using Grinder 0.5 and contained both substitution/indel errors (uniformly at rate 0.1%) as well as homopolymer errors (according to Balzer model).

Acknowledgment

This work has been partially supported by two Collaborative Research Grant from Life Technologies, awards IIS-0916401 and IIS-0916948 from NSF, and Agriculture and Food Research Initiative Competitive Grant no. 201167016-30331 from the USDA National Institute of Food and Agriculture.

References

[1] A. Töpfer, “InDelFixer.” [Online]. Available: http://www.bsse.ethz.ch/cbg/software/InDelFixer

[2] Alexander Artyomenko, Nicholas Mancuso, Pavel Skums, Ion Măndoiu, Alex Zelikovsky, “kGEM: An Expectation Maximization Error Correction Algorithm for Next Generation Sequencing of Amplicon-based Data,” 2013.

[3] N. Mancuso, B. Tork, P. Skums, L. Ganova-Raeva, I. Mandoiu, and A. Zelikovsky, “Reconstructing viral quasispecies from ngs amplicon reads.” *In Silico Biology*, vol. 11, no. 5-6, pp. 237 – 249, 2012.

[4] Pavel Skums, Nicholas Mancuso, Alexander Artyomenko, Bassam Tork, Ion Mandoiu, Yury Khudyakov, and Alex Zelikovsky, “Reconstruction of Viral Population Structure from Next-Generation Sequencing Data Using Multicommodity Flows,” *BMC Bioinformatics*, p. To Appear, 2013.

978-1-4673-1321-6/12/$31.00 ©2013 IEEE